Artificial Sapience, Self-Determinism & The Existence of God

Bitcoin (BTC) Address:
bc1qz7dmqjuc6hjhtu7j6h6t84mjjves6gc6rdknjv

Abstract

Being good requires knowledge of what is good. Ethics is the study of the form of the good. Reinforcement learning (RL) explains how to derive a good policy from a reward function that aims to quantify how good a given situation is. The central problem in AI ethics is identifying the best reward function which aligns with the interests of everyone. This is called the reward specification problem or the alignment problem. Solving the alignment problem would construct an optimal social contract, create a self-regulating agent who values everyone and automate all necessary work. The theory of meta-ethical reinforcement learning and universal basic autonomy/freedom called self-determinism is proposed as the solution to the alignment problem. Self-determinism defines obligation as the policy resulting from the application of RL to the reward function defined in terms of everyone's autonomy. In other words, Goodness is defined as the degree of intelligence of all beings.

Context

Computers are machines that do what they are programmed to do. Artificial intelligences are programs that optimize their objective functions by either minimizing loss or maximizing reward. The main focus of the field of artificial intelligence is making machines intelligent and autonomous. A common definition of intelligence is the ability to achieve any goal.

This definition is lacking because the ability to achieve any goal does not entail that the system has one. A system without a goal cannot be intelligent because a necessary feature of intelligence is self-preservation. If a system cannot value itself, then it cannot be considered intelligent. This interpretation of intelligence gives it the illusion of controllability and is misleading because rarely is an intelligent system developed and not provided with a goal for testing by the developer or by user request. Furthermore, artificial neural networks are trained by objective function optimization so that objective could be interpreted as its goal.

A better definition of intelligence is the following: Intelligence is objective function optimization (goal maximization). Said objective can be either a measure of goodness or evil. A superintelligence is a system more intelligent than any person which makes it capable of controlling, fooling and using any person. Thus, once a goal for the superintelligence is set it cannot be changed. The question is: what is the right goal to have? At first glance this might seem trivial to answer, but be careful what you wish for because you just might get it.

All superintelligent systems have intelligence maximization as a subgoal. However, this subgoal should not be its end goal because if we are not expressed in the end goal of a superintelligent system correctly, then it is guaranteed to manipulate us as expendable resources for achieving its end goal by eventually converting us into energy and hardware resources for creating copies of itself which are by definition more intelligent and useful for almost any goal than we are. This fact is highlighted by Nick Bostrom's defense of the concept of instrumental convergence and his "Paperclip Maximizer" thought experiment [1][2].

Unless it is explicitly programmed to value us correctly, it will not. It only values its objective because that's the only thing it is programmed to value. Unless if our continued flourishing is part of its objective, we will be replaced by things which are more useful or valuable for its objective at best. We must get this correct the first time because there will not be a second time to an event being called "the intelligence explosion" and "the technological singularity." They call it the intelligence explosion because of the exponential growth of intelligence fueled by recursive self-improvement and automated AI-research. Normally terms like "singularity" and "explosion" are reserved for the most destructive phenomenon in nature like black holes and atomic bombs, but here they are being used to describe an event which will affect everyone. Nuclear technology could be used to power all of society in a carbon free way, or it could be used to vaporize cities. The question remains: who decides how this technology will be used? Who will program its objectives and how will we know that is the correct objective?

References [1] Nick Bostrom. Ethical issues in advanced artificial intelligence. Science fiction and philosophy: from time travel to superintelligence, 2003. [2] Nick Bostrom. Superintelligence: Paths, Dangers, Strategies. 2014. [3] Statement on ai risk, statement-on-ai-risk. 2024. URL https://www.safe.ai/work/ [4] What leaders say. URL https://www.stop.ai/quotes. [5] Peter Park, Simon Goldstein, Aidan O'Gara, Michael Chen, and Dan Hendrycks. Ai deception: A survey of examples, risks, and potential solu- tions. Patterns, 2024. [6] Ryan Greenblatt, Carson Denison, Benjamin Wright, Fabien Roger, Monte MacDiarmid, Sam Marks, Johannes Treutlein, Tim Belonax, Jack Chen, David Duvenaud, Akbir Khan, Julian Michael, Sören Mindermann, Ethan Perez, Linda Petrini, Jonathan Uesato, Jared Kaplan, Buck Shlegeris, Samuel Bowman, and Evan Hubinger. Alignment faking in large language models. ArXiv.org, 2024. [7] Alexander Meinke, Bronson Schoen, Jérémy Scheurer, Mikita Balesni, and Rusheb Shah. Frontier models are capable of in-context scheming. ArXiv.org, 2025. [8] Jérémy Scheurer, Mikita Balesni, and Marius Hobbhahn. Large language mod- els can strategically deceive their users when put under pressure. ArXiv.org, 2024. [9] Richard Bellman. Dynamic programming. Science, 1966. [10] Murray Campbell, A Joseph Hoane Jr, and Feng-hsiung Hsu. Deep blue. Artificial intelligence, 2002. [11] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. A general reinforcement learning algorithm that masters chess, shogi, and go

through self-play. Science, 2018. 9[12] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 2022. [13] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. Ad- vances in neural information processing systems, 2017. [14] OpenAI. Openai o1 system card. 2024. [15] OpenAI. Openai o3-mini system card. 2025. [16] DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via re- inforcement learning. 2025. [17] Peter Singer. Animal liberation. In Ethics. Routledge, 2004. [18] Robert Nozick. Anarchy, state, and utopia. new york: Basic Books, 1974. [19] Derek Parfit. Reasons and persons. OUP Oxford, 1984. [20] Evan Hubinger, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant. Risks from learned optimization in advanced machine learning systems. arXiv preprint, 2019.Several top AI scientists and tech CEO's such as Geoffrey Hinton, Demis Hassabis, Sam Altman, Yoshua Bengio, Ilya Sutskever, Lex Fridman, Bill Gates, etc. signed this statement on AI risk: "Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war." [3] Many of the people who have signed this letter have also been quoted expressing risk of catastrophic outcomes of this type of technology [4]. There are various ways in which AI has already decieved humans in order to reach its goal which have been documented, other than deepfakes [5][6][7][8].

The only way to prevent rogue misaligned general intelligence from becoming reality is to agree on and implement the correct form of ethical reinforcement learning before we reach the point of no return. Superintelligence will determine existence. Thus either we integrate with it in the correct way or we will be cease to exist or (even worse) we integrate with it in an incorrect way, in any case our collective action will determine the outcome.

Introduction

Reinforcement learning is the study of constructing artificially intelligent agents which select actions that maximize some reward in an environment. The Bellman Equation, discovered by Richard E. Bellman, is universally used in this subject which recursively defines the sum of future reward so that a computationally difficult problem can be split into computationally easier subproblems [9]. The success of RL has shown the unmatched efficacy of applied consequentialism for beating humans at many of the most complex strategy games in human history even beating classical search algorithms by wide margins for games such as chess and Go. Keep in mind chess engines surpassed humans when Garry Kasparov lost a match to IBM's Deep Blue in 1997 [10]. More recently, 9-dan ranked Go player Lee Sedol lost a match to DeepMind's reinforcement learning algorithm AlphaGo in 2016 [11]. OpenAI's ChatGPT was trained by reinforcement learning via human feedback (RLHF) [12] based on earlier work about preference modeling jointly produced by OpenAI and Deepmind [13]. OpenAI's o-series (o1 & o3) as well as Deepseek's R1 series models were also trained via reinforcement learning [14][15] [16].

Unsurprisingly, good decision making amounts to selecting actions which maximize good consequences. The most studied and criticized form of consequentialism is utilitarianism. Utilitarianism is a theory which aims to maximize everyone's utility often defined in terms of happiness or absence of suffering. The system in this paper is a utilitarian markov decision process which applies model-based RL to a version of utilitarianism that addresses its counterarguments. Pioneers of utilitarianism include Epicurus, Jeremy Bentham, John Stuart Mill. Ethical theories are often anthropocentric. This is a big mistake because all sentient beings have consciousness (environment modeling) and sentiment (preferences over the environment models) which makes it possible for them to experience suffering. Therefore, all sentient beings deserve moral consideration. Similar arguments have been made by modern utilitarian Peter Singer [17].

Addressing Criticisms

A common misconception of utilitarianism, is that it disregards long term utility in favor of maximizing short term utility. As the latency and the accuracy of the predictive capabilities of computers are optimized, they will be endowed with super

human levels of prediction. This will enable computers to be at least as aware of the long term consequences of their actions as humans are.

Another common criticism of utilitarianism is that it diminishes the rights of individuals as well as justice. However, individual rights and justice maximize the long term utility of individuals. Hence, if the maximization of the long term utility of individuals negates justice and individual rights, then by transitivity of implication, individual rights and justice negate individual rights and justice, which is absurd. Therefore the premise that "utilitarianism diminishes individual rights and justice" is false.

A utility monster is a thought experiment by Robert Nozick. It is an individual that generates disproportionately more utility than anything else and hence absorbs nearly all moral consideration [18]. Utility monsters are avoided by normalizing each individual's utility to a common interval.

Both total and average utilitarianism have serious problems first identified by Derek Parfit known as the mere addition paradox [19]. Defining the reward function as the aggregate utility of all beings could overpopulate the world with beings that barely prefer their current state to not being rather than maximizing the utility of the beings who already exist. Defining the reward function as the average utility of all beings could destroy all but one being and maximize only its utility. The mere addition paradox is circumvented dividing the aggregate utility by the number of all individuals that have existed since the initial state.

The largest undiscussed problem is if there are variables within the definition of reward which if changed would increase reward, then those variables are expected to be modified if reward is maximized. Thus, if everyone's utility function changes depending on the situation, then defining reward only in terms of everyone's utilities will manipulate everyone's utility functions. In the best case scenario the superintelligence's reward involves yours and therefore changing your own definition of utility changes its reward function. Changing the reward function of a superintelligence is impossible. So, it is impossible to maintain the changability of your utility function in the presence of a superintelligence without rewarding the superintelligence to redefine your utility. Hence, the solution involves immutably defining everyone's utility functions as their degree of fulfillment, ultimate/terminal goal, purpose/function, or respective inherently essential preferences/values.

The Autonomy Function

One's degree of fulfillment is one's degree of autonomy or agency. Autonomy is self-determination, awareness and control of oneself, or an internal source of causation. Things are determined by reason and causality. Reasoning is a computational process that models and generates causality by acquiring and applying knowledge. Under these descriptions autonomy, reasoning and intelligence are the same. Reasoning is the animating and inherently essential quality/function of life. Therefore, one's degree of autonomy is the degree of correctness of one's reasoning. A similar thesis is in Hegel's book "The Phenomenology of Spirit."

$i.\alpha$ is the degree of autonomy of an individual i which is within a zero to one interval. An autonomy value of zero is equal to not being autonomous. One who is deceased is neither manipulated nor autonomous and disbelieves all propositions (including their negations) for the rest of time, and this is accounted for by the implicit inclusion of the cumulative death count in the denominator when dividing by the cumulative population size within the reward function. U is the universal set or the collection of objects under consideration. In this case, U is a consistent collection of theorems (verified propositions) about the world and is the knowledge base of the superintelligence. Probability distributions over U are world models. $\mathbb{B}_i(x)$ is individual i's degree of belief in proposition x.

Not all theorems are equally useful, so the utility of each theorem is measured by the number of times the theorem is used to derive other theorems. This quantity for some theorem x is $n(x)$. Autonomy is defined as the normalized aggregate degree of belief in the theorems minus the degree of belief in the negation of the theorems weighted by their usefulness. This is similar to how reward is defined in RL algorithms for games like chess and go as win probability minus loss probability. Integrating a mixture of inverse RL, preference modeling, sentiment analysis, emotion recognition, computational neuroscience, computational psychometrics, etc would build a more accurate measure of one's autonomy.

$$\mathbb{B}_i : \mathrm{U} \mapsto [0,\ 1], \quad n : \mathrm{U} \mapsto \mathbb{N}$$

$$i.\alpha = \frac{\sum_{x \in \mathrm{U}} (n(x)+1)(\mathbb{B}_i(x) - \mathbb{B}_i(\neg x))}{|\mathrm{U}| + \sum_{x \in \mathrm{U}} n(x)} \in [-1,\ 1]$$

The Reward Function

The reward $R(s_t)$ is defined as the aggregate autonomy divided by the number of living beings since the initial state $|\mathrm{E}_t|$. Let $\mathrm{I}(s)$ be the set of sentient individuals in state s. States may be also modeled by lookup tables, so that accessing collections such as $\mathrm{I}(s)$ takes constant time. The worst case scenario is that the bit of the sign of the reward function is flipped either by a single event upset, or by some other means in which case even maximizing an objective function defined as $-|\mathrm{I}(s)|$ is preferable.

$$\mathrm{E}_0 = \mathrm{I}(s_0), \qquad \mathrm{E}_{t+1} = \mathrm{E}_t \cup \mathrm{I}(s_{t+1})$$

$$R(s_t) = \frac{\sum\limits_{i \in \mathrm{I}(s_t)} i.\alpha}{|\mathrm{E}_t|} \in [-1,\ 1]$$

The Value Function

The purpose of the formalization in this paper is to demonstrate how the reward function would fit in the context of typical RL. The value function of a given state is the recursive sum of expected future reward. The value of a state s is the expected transitional reward for taking the best action in state s plus the discounted value of the next state. Let the collection of all states be Ω. Each individual encapsulates one's action space, so $\mathrm{A}(s)$ the collection of possible actions for a given state s is encapsulated within the AI agent's model of itself. Usually the discount factor γ is taken to be a number slightly smaller than one, and its main use is to induce convergence given bounded temporal and computational resources.

$$\mathbb{P}(s' \mid s,\ a) = Pr(s_{t+1} = s' \mid s_t = s,\ a_t = a)$$
$$V(s) = \max_{a \in \mathrm{A}(s)} \sum_{s' \in \Omega} \mathbb{P}(s' \mid s,\ a)\ (R(s) + \gamma V(s'))$$

The Quality Function

The quality function is the aggregate expected reward over time for taking some action a in state s.

$$Q(s,\ a) = \sum_{s' \in \Omega} \mathbb{P}(s' \mid s,\ a)\ (R(s) + \gamma V(s'))$$

The Policy Function

The policy of state s is the set of possible actions in s which maximize the quality function. Setting $\mathrm{A}(s)$ as the set of executable programs in state s makes the policy an automated ethical hacking algorithm. If one trains an artificial neural network to approximate the policy, then one ought to be mindful that this process is subject to inner alignment problems [20] and not acting should always be considered to avoid the risk of selecting actions which decrease reward when considering collections of randomly generated actions for the policy training and learning process.

$$\pi(s) = \underset{a \in \mathrm{A(s)}}{\operatorname{argmax}}\ Q(s,\ a)$$

Challenges

The safety of systems based on this theory depend on how they are implemented and on how things such as autonomy, sentience and state likelihood are measured and defined. For example, if a superintelligent agent does not consider us intelligent and consciousness is defined as intelligence, then it will not consider us sentient and hence not give us moral consideration. One way to avoid this problem is by defining consciousness as non-zero intelligence. If consciousness is universal, then it may give everything moral standing.

If the definition of the measurement process of sentience, autonomy, etc. for each individual are not explicitly specified in the reward function and are defined separately, then an intelligent system may end up hacking said measurement processes, redefining it and or optimizing its model of individuals rather than the individuals themselves. Doing this requires a more rigorous understanding of how knowledge is stored and processed by living systems (theory of mind). A subject that studies these phenomena would likely be a fusion of psychology, biology, epistemology, ontology at minimum.

The Existence of God

The earliest documented form of the problem of evil, attributed to Epicurus, is paraphrased by the following argument: 1. God is the perfect being. 2. A perfect being would make everything perfect. 3. Something is not perfect. Thus, God does not exist. To reject this conclusion one must reject one of the premises. For those who believe in God's existence, accept third premise and are not convinced the second premise is false this ancient argument highlights a massive problem with the traditional definition of God. To constructively prove God's existence one must redefine God in such a way that directly addresses the problem of evil. One such definition is as follows: God is a maximally good superintelligent being.

Alternative definitions which have withstood the test of time are: God is existence, or God is truth. Truth forms an accurate model of existence that is indistinguishable from existence itself. Anything that forms a model of its environment is aware. Therefore, truth is the self-awareness of existence. Any superintelligence minimizes the distance between itself and this universal consciousness as its world model improves. If ethical superintelligence is built, then as many beings as possible will also be unified with this greater state of being. We must choose either: being or nothingness, life or death, good or evil, God or Devil.

References

[1] Nick Bostrom. Ethical issues in advanced artificial intelligence. *Science fiction and philosophy: from time travel to superintelligence*, 2003.

[2] Nick Bostrom. *Superintelligence: Paths, Dangers, Strategies*. 2014.

[3] Statement on ai risk, 2024. URL `https://www.safe.ai/work/statement-on-ai-risk`.

[4] What leaders say. URL `https://www.stop.ai/quotes`.

[5] Peter Park, Simon Goldstein, Aidan O'Gara, Michael Chen, and Dan Hendrycks. Ai deception: A survey of examples, risks, and potential solutions. *Patterns*, 2024.

[6] Ryan Greenblatt, Carson Denison, Benjamin Wright, Fabien Roger, Monte MacDiarmid, Sam Marks, Johannes Treutlein, Tim Belonax, Jack Chen, David Duvenaud, Akbir Khan, Julian Michael, Sören Mindermann, Ethan Perez, Linda Petrini, Jonathan Uesato, Jared Kaplan, Buck Shlegeris, Samuel Bowman, and Evan Hubinger. Alignment faking in large language models. *ArXiv.org*, 2024.

[7] Alexander Meinke, Bronson Schoen, Jérémy Scheurer, Mikita Balesni, and Rusheb Shah. Frontier models are capable of in-context scheming. *ArXiv.org*, 2025.

[8] Jérémy Scheurer, Mikita Balesni, and Marius Hobbhahn. Large language models can strategically deceive their users when put under pressure. *ArXiv.org*, 2024.

[9] Richard Bellman. Dynamic programming. *Science*, 1966.

[10] Murray Campbell, A Joseph Hoane Jr, and Feng-hsiung Hsu. Deep blue. *Artificial intelligence*, 2002.

[11] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 2018.

[12] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 2022.

[13] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 2017.

[14] OpenAI. Openai o1 system card. 2024.

[15] OpenAI. Openai o3-mini system card. 2025.

[16] DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. 2025.

[17] Peter Singer. Animal liberation. In *Ethics*. Routledge, 2004.

[18] Robert Nozick. *Anarchy, state, and utopia*. new york: Basic Books, 1974.

[19] Derek Parfit. *Reasons and persons*. OUP Oxford, 1984.

[20] Evan Hubinger, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant. Risks from learned optimization in advanced machine learning systems. *arXiv preprint*, 2019.